UMA METODOLOGIA PARA IDENTIFICAÇÃO DE MÓDULOS FORMADORES DE SEQUÊNCIAS DE PROTEÍNAS MOSAICAS DO *Trypanosoma cruzi* A PARTIR DO TRANSCRIPTOMA DO PARASITO UTILIZANDO A FERRAMENTA BLAST

Elisa Boari de Lima[1], André de Souza Gomes[1], Thiago de Souza Rodrigues[2]

Universidade Federal de Lavras
Departamento de Ciência da Computação
C. P. 3037 – Campus da UFLA – 37200-000 Lavras – MG – Brasil

[1](elisa, andregomes)@comp.ufla.br, [2]thiago@dcc.ufla.br

Resumo. Este trabalho propôs uma metodologia de identificação de módulos formadores de sequências nucleotídicas codificadoras de proteínas mosaicas do *Trypanosoma cruzi* utilizando a ferramenta BLAST. Para o desenvolvimento da metodologia, foi utilizada a família MASP de proteínas e aplicado inicialmente o conjunto de valores padrão dos parâmetros da ferramenta. Posteriormente foram estudadas diferentes combinações de valores de parâmetros a fim de comparação de resultados, incluindo valores indicados pela literatura. A metodologia desenvolvida provou ser eficaz para o objetivo proposto, obtendo melhores resultados quando aplicados valores diferentes dos valores padrão para filtro de regiões de baixa complexidade, *E-value* e tamanho inicial de palavra.

Palavras-chave: Bioinformática, Proteínas Mosaicas, *Trypanosoma cruzi*, Transcriptoma, BLAST.

1. Introdução

O *T. cruzi* é um protozoário causador da doença de Chagas, uma doença incurável e debilitante que afeta milhões de pessoas nas Américas Central e Latina. O sequenciamento do genoma desse parasito permitiu o início de análises de sequências de nucleotídeos e aminoácidos derivadas a fim de identificar diversos novos alvos terapêuticos potenciais, além de fornecer dados estruturais para estudos funcionais posteriores como os dados sobre módulos encontrados em determinadas proteínas formadas pelo rearranjo genético desses e conhecidas como proteínas mosaicas.

Um módulo pode ser definido como um conjunto de aminoácidos invariáveis ou altamente conservados usado repetidamente como "blocos de construção" em diversas proteínas. Cada módulo pode apresentar uma função enzimática, sinalizadora, regulatória ou estrutural diferente, o que faz com que a arquitetura modular de proteínas permita a evolução dessas com funções complexas e altamente especializadas.

A grande variabilidade clínica e epidemiológica da doença de Chagas, associada às características genéticas da população do *T. cruzi*, torna o tratamento da doença limitado a medicamentos utilizados desde o final da década de 1960, apresentando taxas de efeitos colaterais altas e eficácia variável durante a fase crônica da doença. Por essa razão, a identificação dos módulos constituintes das proteínas de uma das famílias protéicas necessárias à sobrevida e à patogenicidade do parasito (a família MASP) por meio da análise de seu transcriptoma abre caminho para a busca de novas estratégias terapêuticas e para a identificação de novos biomarcadores importantes para o desenvolvimento de novas drogas e prognóstico clínico da doença de Chagas.

Os objetivos gerais deste trabalho foram o desenvolvimento de uma metodologia para identificação dos módulos formadores de sequências nucleotídicas codificadoras de proteínas mosaicas e, dada uma família de proteínas do *T. cruzi*, a verificação de que essas apresentam estrutura mosaica, ou seja, são formadas por módulos que se repetem em diferentes proteínas da família. Para o desenvolvimento da metodologia foi utilizada a família MASP de proteínas do *T. cruzi*.

2. Expressão Genômica

Todo organismo possui um genoma que contém a informação biológica necessária para construir e manter um exemplar vivo. O genoma é um deposito de informação biológica, mas sozinho é incapaz de liberar tal informação para a célula. A utilização da informação biológica requer uma atividade coordenada de enzimas e outras proteínas, que participam em uma série complexa de reações bioquímicas chamada expressão genômica [1].

O produto inicial da expressão genômica é o transcriptoma, uma coleção de moléculas de RNA derivadas dos genes codificadores de proteínas cuja informação biológica é requisitada pela célula em um dado momento. Essas moléculas de RNA direcionam a síntese do produto final da expressão genômica, o

proteoma, o repertório celular de proteínas, que especificam a natureza das reações bioquímicas que a célula é capaz de realizar. O transcriptoma é construído pelo processo chamado transcrição, no qual genes individuais são copiados para moléculas de RNA. A construção do proteoma envolve a tradução dessas moléculas de RNA em proteína [1].

Sequências de RNAm espelham a sequência de DNA dos genes dos quais foram transcritas. Consequentemente, por meio da análise do transcriptoma, pesquisadores podem determinar o momento e o local em que um gene é ativado ou desativado em vários tipos de células e tecidos. A depender da técnica utilizada é possível contar o número de transcrições para determinar a quantidade de atividade genética, ou nível de expressão, em certo tipo de célula ou tecido. Quanto maior o número de transcrições, geralmente mais importante aquela transcrição é para o funcionamento celular [2]. Estudos com transcriptoma ajudam a explicar uma sequência de genoma, dando apoio à identificação dos genes cujos papéis no genoma não foram determinados por outros métodos [1].

3. Proteínas Mosaicas

Segundo [3], proteínas mosaicas são um grupo de proteínas que podem ser formadas por um ou mais tipos de uma variedade de diferentes módulos estruturais e que possuem uma extensão diversa de funções. A Figura 1 mostra a representação gráfica da estrutura de proteínas mosaicas.

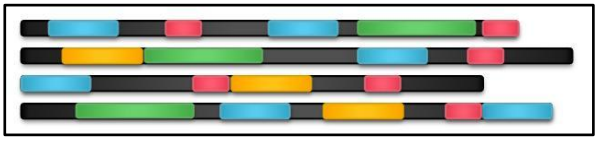

Figura 1 – Representação Gráfica da Estrutura de Proteínas Mosaicas

De acordo com [4], a análise comparativa de sequências protéicas tem revelado que muitas proteínas extracelulares são constituídas por um repertório limitado de padrões ou módulos de sequência. Tais proteínas, chamadas proteínas mosaicas, podem então ser descritas como a justaposição linear de módulos contíguos. Módulos podem ser definidos como subsequências usadas repetidamente como "blocos de construção" em diversas sequências protéicas e provavelmente têm aparecido por meio da "mistura" de genes [5]. Várias proteínas mosaicas possuem papel essencial na série de reações químicas da biologia extracelular [4].

Conforme [6], muitas proteínas são compostas por um número de domínios discretos, que frequentemente estão envolvidos em funções específicas que contribuem para a atividade geral da proteína. Uma análise dos genes codificadores de proteínas mosaicas revela uma forte correlação entre organização de domínio e estrutura *intron-exon*. Em outras palavras, cada domínio tende a estar codificado por um ou uma combinação de *exons* que iniciam e terminam no mesmo quadro de *splice*. Proteínas mosaicas aparentam ser criadas pela junção de múltiplos domínios por meio do embaralhamento de *exons.*

Os domínios encontrados em proteínas mosaicas são evolucionariamente móveis, o que significa que eles se espalharam durante a evolução e agora ocorrem em proteínas que antes não estariam relacionadas [7]. A maioria das proteínas mosaicas é extracelular ou constitui as partes extracelulares de proteínas ligadas a membrana, por isso foi proposto que proteínas mosaicas desempenharam um importante papel na evolução da multicelularidade [8].

4. *Trypanosoma cruzi*

O *T. cruzi* infecta e se adapta ao hospedeiro vertebrado explorando estratégias evolucionárias para invadir células alvo e evadir (ou confundir) o sistema imunológico [9]. O processo de invasão, evasão e infecção envolve diferentes famílias de proteínas de superfície [10]. Uma estratégia chave é a geração e apresentação de antígenos de superfície variáveis [11]. O parasito pode tirar vantagem dessa estratégia para aderir a diferentes moléculas na membrana celular e matriz extracelular da célula hospedeira [10].

O sequenciamento do genoma do *T. cruzi* foi concluído em 2005, mas devido às muitas dificuldades com sequências repetitivas e heterozigose do clone, o sequenciamento foi apenas parcial. Foram preditas 22.570 proteínas, das quais 12.570 formam pares alélicos [12].

Em termos biológicos, o *T. cruzi* apresenta características bastante peculiares que se refletem na organização e função de seu genoma. As sequências repetitivas do DNA do *T. cruzi* representam pelo menos 50% de todo o seu genoma e são formadas principalmente pelas famílias de genes que compõem as proteínas de superfície. Esses totalizam 18% dos genes codificadores de proteínas do *T. cruzi*. A família MASP do *T. cruzi*, utilizada neste trabalho, é uma família de proteínas de superfície associadas à mucina que contém 1.377 membros, o que corresponde a aproximadamente 6% do genoma diplóide do *T. cruzi*, e é caracterizada por regiões centrais altamente variáveis e que frequentemente contêm sequências repetidas [13].

O baixo número de peptídeos detectados por abordagens proteômicas sugere que proteínas da família MASP podem conter extensivas modificações após o processo de tradução. Genes da família MASP podem ser expressos em estágios intermediários não representados nos dados do proteoma ou podem ser expressos de modo mutuamente exclusivo [13].

5. Alinhamento de Sequências por Pares

Alinhamento de sequências é o procedimento de se comparar duas (alinhamento por pares) ou mais (alinhamento múltiplo) sequências de ácidos nucléicos (DNA e RNA) ou proteína por meio da busca de uma série de caracteres individuais ou padrões de caracteres que estão na mesma ordem nas sequências [14].

O alinhamento de sequências busca possibilitar ao pesquisador determinar se duas sequências apresentam similaridade suficiente tal que uma inferência sobre homologia possa ser justificada. Homologia significa que duas ou mais sequências têm um ancestral comum. Já similaridade, que é um forte argumento para homologia, é uma medida da qualidade do alinhamento entre duas sequências com base em algum critério. A similaridade não se refere a nenhum processo histórico, sendo apenas uma comparação das sequências com algum método, podendo ser definida, por exemplo, contanto posições idênticas entre duas sequências.

Há duas formas de alinhamento por pares: global e local. No alinhamento global, é feita uma tentativa de alinhar toda a extensão das sequências. Sequências que são bastante semelhantes e que possuem aproximadamente o mesmo tamanho são candidatas ao alinhamento global. No alinhamento local, são alinhadas extensões de sequência com alta densidade de casamentos, gerando desse modo uma ou mais ilhas de casamentos ou sub-alinhamentos nas sequências alinhadas. Alinhamentos locais são mais apropriados para sequências que são semelhantes apenas em partes de suas extensões, sequências com tamanhos diferentes ou sequências que compartilham um domínio ou região conservada [14]. A Figura 2 exemplifica a diferença entre os dois tipos de alinhamento.

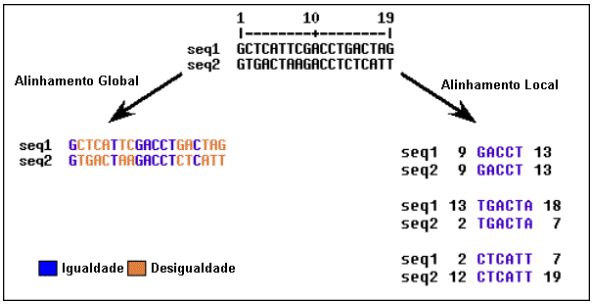

Figura 2 – Exemplo de Alinhamento Global (a) e Local (b)

O algoritmo Smith-Waterman é utilizado para produzir alinhamentos locais entre pares de sequências de aminoácidos ou nucleotídeos. Em um alinhamento local, o alinhamento é interrompido nas extremidades de regiões de alta similaridade e é dada uma prioridade muito maior à busca de tais regiões do que à extensão do alinhamento para incluir mais pares de aminoácidos ou nucleotídeos vizinhos. Esse tipo de alinhamento favorece a localização de padrões conservados de nucleotídeos em sequências de ácidos nucléicos ou de domínios de aminoácidos em sequências protéicas.

6. BLAST: *Basic Local Alignment Search Tool*

O BLAST, uma das principais ferramentas da bioinformática, utiliza um método heurístico que se baseia na determinação de trechos de similaridade local por meio da comparação de sequências de ácidos nucléicos ou proteínas contra sequências armazenadas em um banco de dados, calculando a significância estatística para os resultados obtidos após as comparações. Essa ferramenta pode ser utilizada para inferência de relações funcionais evolutivas de várias sequências, assim como para auxiliar na identificação de membros de famílias gênicas.

O algoritmo do BLAST aumenta a velocidade do alinhamento de sequências buscando primeiro por palavras ou *k*-tuplas comuns à sequência buscada (*query*) e a cada sequência de um banco de dados. A busca é delimitada às palavras mais significativas. No algoritmo do BLAST, o tamanho da palavra é, por padrão, três para proteínas e 11 para ácidos nucléicos. Esses tamanhos são o mínimo necessário para alcançar uma pontuação de palavra alta o suficiente para ser significativa, mas não tão elevada que padrões curtos embora significativos sejam perdidos [14].

O BLAST é basicamente um conjunto de programas que buscam em bancos de dados de sequências por similaridades estatisticamente significativas. Essa busca precisa de vários passos e parâmetros de controle. Os cinco programas tradicionais do BLAST são: BLASTN, BLASTP, BLASTX, TBLASTN e TBLASTX. Os quatro últimos realizam comparação de sequências protéicas, enquanto o BLASTN trabalha com comparação de sequências de ácidos nucléicos [15]. Neste trabalho foi utilizado o programa BLASTN.

6.1. BLASTN

O BLASTN tem como entrada uma sequência de nucleotídeos e a compara com um banco de dados de ácidos nucléicos. Esse programa é muito utilizado para procurar sequências que são muito conservadas, sendo tipicamente aplicado para classificação de elementos repetitivos, exploração de sequências entre espécies, explicação de DNA genômico e clusterização de estudos protéicos.

Como a molécula de DNA tem fita dupla e genes podem ocorrer em ambas as fitas, quando uma sequência *query* é comparada com um banco de dados o BLASTN examina ambas as suas fitas: a sequência original com rótulo positivo e seu complemento reverso, com rótulo negativo. Como o BLAST alinha apenas caracteres e não tem nenhum modelo de genes ou outras características incluídas, é impossível

determinar a partir de um alinhamento BLASTN em que fita o gene está [15].

6.2. Parâmetros do BLAST

O algoritmo do BLAST é controlado por uma série de parâmetros, muitos dos quais possuem valores padrão e não precisam ser explicitamente determinados. Os parâmetros do BLASTN utilizados neste trabalho e tidos como mais relevantes são detalhados abaixo.

Filtro de Regiões de Baixa Complexidade – a filtragem pode eliminar informações estatisticamente significantes porém biologicamente desinteressantes do relatório do BLAST, deixando apenas as regiões biologicamente mais interessantes da sequência *query* disponíveis para casamento específico contra as sequências do banco de dados. O BLAST usa por padrão a filtragem DUST para o BLASTN e SEG para os outros programas [16].

***E-value* –** indica a validade de um alinhamento por avaliar a probabilidade de ele ocorrer por acaso. Quanto menor o valor, mais provável de ser um bom alinhamento e representar uma similaridade real ao invés de um alinhamento aleatório [16]. Por padrão, o BLAST mostra alinhamentos com valores de *E-value* de no máximo dez.

Tamanho Inicial de Palavra – é um dos parâmetros mais importantes que dirigem a sensibilidade de buscas BLAST. Os valores padrão são três para sequências de proteínas e 11 para sequências de ácidos nucléicos.

Sistema de Pontuação para Nucleotídeos – buscas de nucleotídeos usam um sistema de pontuação simples que consiste em uma "recompensa" para igualdade de nucleotídeos e uma "penalização" para desigualdade. A razão recompensa/penalização absoluta deve ser aumentada à medida que a divergência de sequências aumenta [16]. O BLAST usa por padrão o sistema 1/-3.

6.3. Busca por Casamentos Curtos

Sequências curtas (menos de 20 nucleotídeos) frequentemente não encontrarão casamentos significativos com as entradas do banco de dados sob as configurações padrão do BLAST. As razões gerais para isso são que o limiar de significância definido pelo *E-value* é estabelecido muito rigorosamente e o tamanho de palavra padrão é definido muito alto. Esses parâmetros devem ser ajustados para trabalhar com sequências curtas. O filtro de baixa complexidade também é removido visto que elimina porcentagens maiores de uma sequência curta, podendo até mesmo eliminar a *query* [17].

Buscas com sequências curtas podem ser praticamente idênticas e apresentar um *E-value* relativamente alto. Isso se deve ao fato de que o cálculo do *E-value* leva em consideração o tamanho da sequência *query* e ao fato de que sequências curtas têm uma alta probabilidade de ocorrer no banco de dados puramente ao acaso. Essa é a razão pela qual os *E-values* são definidos em valores muito altos quando executando buscas no BLAST usando sequências curtas tanto de nucleotídeos quanto de aminoácidos [16].

A Tabela 1 apresenta um conjunto de parâmetros sugerido por [18] para buscas com sequências de nucleotídeos curtas.

Parâmetro	Valor Padrão	Valor Indicado
Tamanho de Palavra	11	7
E-value	10	1000
Filtro de Complexidade	Ativado (T)	Desativado (F)

Tabela 1 – Parâmetros do BLAST para Sequências Nucleotídicas Curtas

7. Metodologia

Os dados utilizados neste trabalho são sequências de nucleotídeos codificadoras de proteínas da família MASP do *T. cruzi* constituintes do transcriptoma do parasito. As 810 sequências estudadas, organizadas em um arquivo no formato FASTA (entrada padrão para o BLAST), foram obtidas junto ao Departamento de Parasitologia do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, que tem o *T. cruzi* como uma de suas linhas de pesquisa na subárea de Protozologia.

Como ainda não foi proposta na literatura uma metodologia para identificação de módulos formadores de proteínas mosaicas, neste trabalho foram empregados procedimentos para criação de uma estratégia que possibilitasse o desenvolvimento da metodologia proposta de identificação de módulos formadores de proteínas mosaicas do T. cruzi a partir do transcriptoma do parasito. Este trabalho foi realizado em paralelo com o trabalho de [19], que propõe uma metodologia de identificação de módulos de proteínas mosaicas do T. cruzi a partir do proteoma do parasito.

Este trabalho utilizou a ferramenta BLAST versão 2.2.17, em especial os programas *formatdb* e BLASTN para criação de bancos de dados a partir de arquivos contendo sequências em formato FASTA e para realização de alinhamentos entre pares de sequências nucleotídicas, respectivamente.

Antes que pudesse ser desenvolvida a metodologia de identificação de módulos buscou-se um modo de encontrar tais módulos na família MASP do *T. cruzi* por meio do alinhamento inicial de todas as sequências contra todas e da análise dos resultados obtidos. O BLAST inicialmente foi executado para o conjunto S_1 das 810 sequências da família MASP com

os valores padrão dos parâmetros gerando o alinhamento A_1. Os resultados foram filtrados segundo a estratégia detalhada na Seção 7.1, mantendo-se assim apenas os alinhamentos considerados mais significativos.

O resultado do BLAST possibilitou verificar visualmente a existência de regiões comuns a várias sequências da família. A Figura 3 traz um exemplo, mostrando alinhamentos da mesma região da sequência *query* com três sequências distintas do banco de dados. Considerando a estrutura das proteínas como modular, tais módulos estariam presentes nos alinhamentos filtrados. Foi então realizada a extração das regiões envolvidas em tais alinhamentos para que pudesse ser feita uma análise específica das mesmas, o que exigiu a criação de uma estratégia de corte, detalhada na Seção 7.1. As regiões extraídas compõem o conjunto S_2.

```
>Tc00.1047053506703.20
          Length = 1242

 Score =  127 bits (64), Expect = 4e-030
 Identities = 64/64 (100%)
 Strand = Plus / Plus

Query: 1  atggcgatgatgatgactggccgtgtgctgctggtgtgtgccctctgcgtgctgtggtgc 60
          ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1  atggcgatgatgatgactggccgtgtgctgctggtgtgtgccctctgcgtgctgtggtgc 60

Query: 61 ggtg 64
          ||||
Sbjct: 61 ggtg 64


>Tc00.1047053511603.190
          Length = 879

 Score =  127 bits (64), Expect = 4e-030
 Identities = 64/64 (100%)
 Strand = Plus / Plus

Query: 1  atggcgatgatgatgactggccgtgtgctgctggtgtgtgccctctgcgtgctgtggtgc 60
          ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1  atggcgatgatgatgactggccgtgtgctgctggtgtgtgccctctgcgtgctgtggtgc 60

Query: 61 ggtg 64
          ||||
Sbjct: 61 ggtg 64


>Tc00.1047053511089.30
          Length = 759

 Score =  115 bits (58), Expect = 2e-026
 Identities = 58/58 (100%)
 Strand = Plus / Plus

Query: 7  atgatgatgactggccgtgtgctgctggtgtgtgccctctgcgtgctgtggtgcggtg 64
          ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1  atgatgatgactggccgtgtgctgctggtgtgtgccctctgcgtgctgtggtgcggtg 58
```

Figura 3 – Alinhamento de uma mesma região da *query* com três sequências distintas do banco de dados.

Observou-se em A_1 a existência de transitividade entre os alinhamentos filtrados, ou seja, considerando, por exemplo, regiões A, B e C de sequências, se A alinhou com B, que por usa vez alinhou com C, então A também alinhou com C. Devido a esse fato a ocorrência de redundância é bastante evidente, sendo necessária sua eliminação. Para isso, alinhou-se todas as sequências S_2 contra elas mesmas realizando, posteriormente, a filtragem dos resultados. Com isso, criou-se uma estratégia de separação dos resultados em grupos (Seção 7.1), de modo a se ter em cada grupo apenas sequências semelhantes, ou seja, que apresentam partes idênticas entre si. A separação dos grupos possibilitou observar que uma mesma sequência pode estar em diferentes grupos, visto que pode conter mais de uma região comum às demais sequências. Aplicou-se a estratégia de corte aos grupos para se obter uma subsequência representativa para cada um, gerando o conjunto S_3. Desse modo conseguiu-se reduzir o número de sequências presentes em mais de um grupo.

Para chegar aos possíveis módulos, notou-se a necessidade de adoção de uma abordagem iterativa dos passos descritos anteriormente até que cada grupo contenha apenas uma sequência: um possível módulo "escolhido" como representante ao longo do processo. A cada iteração a estratégia de corte define um representante de cada grupo para ser colocado no novo arquivo gerado, sendo que os tamanhos dos grupos vão sendo reduzidos até se tornarem unitários, ou seja, até que a sequência de um grupo só alinhe com ela mesma, sendo assim uma candidata a módulo que representa todas as outras sequências dos grupos dos quais ela participou ao longo do processo. Essa é a idéia da metodologia desenvolvida neste trabalho.

Criou-se uma última estratégia para definição de quais módulos candidatos seriam considerados módulos da família de proteínas mosaicas (Seção 7.1).

7.1. Estratégias Utilizadas

As estratégias criadas e utilizadas neste trabalho são descritas a seguir.

Formatação do Relatório do BLAST

O relatório do BLAST apresenta diversas informações sobre todos os alinhamentos significativos encontrados. Para a metodologia desenvolvida, as informações relevantes ao processamento são apenas a identificação das sequências envolvidas no alinhamento, as posições alinhadas em cada uma e o valor de identidade. A rotina de formatação do relatório extrai essas informações organizando-as de modo a facilitar sua manipulação, como mostra a Figura 4, em que cada linha representa um alinhamento e as colunas representam, respectivamente, a identificação da sequência *query* e da sequência do banco de dados, as posições iniciais e finais dos alinhamentos na *query* e na sequência do banco de dados e a porcentagem de identidade.

Tc00.1047053408547.10	Tc00.1047053506879.10	339	351	290	302	92
Tc00.1047053408547.10	Tc00.1047053508081.30	339	351	287	299	100
Tc00.1047053508869.59	Tc00.1047053510625.54	4	21	3	20	94
Tc00.1047053508869.59	Tc00.1047053508481.20	1	19	1	19	94
Tc00.1047053507071.349	Tc00.1047053510039.30	1	28	3	30	50

Figura 4 – Exemplo de Formatação do Relatório do BLAST.

Filtragem dos Resultados

Esta estratégia elimina dos resultados aqueles alinhamentos que não obedecem a um determinado patamar de identidade, conservando apenas os alinhamentos considerados significativos. Foi definido que para sequências nucleotídicas apenas os

alinhamentos com 100% de identidade seriam considerados, visto que a variação de apenas um nucleotídeo leva à modificação do códon, que por sua vez pode codificar um aminoácido diferente, podendo assim modificar alguma característica da proteína codificada pela sequência nucleotídica.

Estratégia de Corte

Esta estratégia é utilizada em duas situações: é aplicada ao alinhamento das sequências originais e aos grupos criados nas iterações. Considerando os alinhamentos de uma dada região em uma dada sequência *query*, a idéia central dessa estratégia é encontrar a subsequência da *query* que está presente na maioria dos alinhamentos, senão em todos. Para isso se estuda as posições alinhadas da *query* em cada um desses alinhamentos e se usa como posições de corte aquelas que mais se repetem entre as posições iniciais e finais dos alinhamentos.

Definidas as posições de corte inicial e final, essas são avaliadas para verificar se limitam uma subsequência com o tamanho mínimo estipulado, que, para sequências nucleotídicas, foi definido como 12 nucleotídeos. Esse tamanho, correspondente à codificação de quatro aminoácidos, foi estabelecido para evitar que se obtenham módulos a partir de ocorrências aleatórias, o que pode ocorrer se forem considerados tamanhos menores. Caso a subsequência definida pelas posições de corte obedeça a essa restrição de tamanho, ela é inserida no novo conjunto S_i de sequências sendo construído. Caso contrário, as posições de corte são descartadas e busca-se um novo par que limite uma subsequência que obedeça à restrição de tamanho. Na ausência de tal par de posições, a região sendo trabalhada é descartada.

Separação de Grupos

O objetivo desta estratégia é agrupar sequências similares. Considerando as sequências A, B, C, D, E e F, os alinhamentos A-B, A-C, B-C, C-D, C-F e E-F, a separação em grupos é feita da seguinte forma: como inicialmente ainda não foram criados grupos, cria-se um novo grupo G_1 do qual A é cabeça; todas as sequências com as quais A se alinha são inseridas no mesmo grupo. Para o exemplo, neste ponto G_1 = {A, B, C}.

Passa-se então aos alinhamentos com a *query* B. Inicialmente se busca todos os grupos a que B pertence (G_1 no caso do exemplo). Cada alinhamento de B é analisado a fim de se verificar se a sequência com que B se alinha pertence a algum grupo ao qual B pertence. Caso a sequência não esteja em nenhum grupo de B, ela é inserida no grupo em que B for cabeça; se tal grupo não existir, cria-se um novo grupo em que B é cabeça e insere-se a sequência nesse novo grupo. Para o exemplo, verifica-se que C já pertence a G_1.

Continuando o processo passa-se a analisar os alinhamentos da *query* C. O mesmo processo para B é realizado. Para o exemplo, ao analisar os alinhamentos de C verifica-se que D não pertence a G_1. Como C ainda não é cabeça de grupo, é criado um novo grupo G_2 = {C, D} e se passa ao próximo alinhamento (C-F). Como F não está em nenhum grupo a que C pertence mas C é cabeça do grupo G_2, F é inserido em G_2, que passa a ser G_2 = {C, D, F}.

A análise continua com os alinhamentos de E, visto que no exemplo não há alinhamentos cuja *query* é D. Neste ponto atinge-se um novo caso: E não está em nenhum grupo, portanto é criado um novo grupo G_3 = {E, F}.

A Tabela 2 apresenta a configuração final dos grupos para o exemplo dado.

G_1	G_2	G_3
A	C	E
B	D	F
C	F	

Tabela 2 – Exemplo de Separação em Grupos

A estratégia de corte será aplicada aos grupos e definirá uma subsequência da cabeça de cada um como representante de todo o grupo. Para o exemplo dado, as posições de corte que definem a subsequência de A representante de G_1 serão definidas com base nas posições dos alinhamentos A-B e A-C; a representante de G_2, com base nas posições de C-D e C-F; e a representante de G_3, com base nas posições do alinhamento E-F.

Definição dos Módulos

Definidos os possíveis módulos ao fim do algoritmo iterativo, realiza-se o alinhamento desses com as sequências originais. São considerados módulos aqueles módulos candidatos que alinham toda a sua extensão com 100% de identidade com pelo menos 1% das sequências. Essa porcentagem mínima de alinhamentos foi definida considerando que candidatos presentes em menos de 1% das sequências da família têm grande chance de ocorrer ao acaso.

7.2. Valores de Parâmetros do BLAST Utilizados

Buscou-se na literatura os valores de parâmetros mais indicados para se trabalhar com sequências curtas, o que foi descrito na Seção 6.3. Os valores utilizados para comparação de resultados da metodologia desenvolvida aplicada à família MASP do *T. cruzi* são apresentados na Tabela 3. A coluna "Valor Indicado" corresponde aos valores para sequências nucleotídicas curtas indicadas por [18]. A coluna "Outros Valores" corresponde a valores testados diferentes dos valores padrão e dos indicados pela literatura. O sistema de pontuação de nucleotídeos é apresentado no formato (recompensa/penalidade).

Parâmetro	Valor Padrão	Valor Indicado	Outros Valores
E-value	10	1000	500
Filtro de Baixa Complexidade	Ativado (T)	Desativado (F)	---
Tamanho Inicial de Palavra	11	7	8-10, 12, 13
Sistema de Pontuação	(1/-3)	---	(1/-2), (1/-1)

Tabela 3 – Valores de Parâmetros Utilizados para Comparação de Resultados

O *E-value* 500 foi escolhido para teste por ser um valor intermediário entre os valores padrão e indicado pela literatura. Os sistemas de pontuação (1/-2) e (1/-1) foram testados por serem indicados para sequências 95% e 75% conservadas, respectivamente, segundo [16].

7.3. Metodologia Desenvolvida

O processo descrito para chegar à metodologia de identificação de módulos formadores de proteínas mosaicas proposta utilizou valores padrão dos parâmetros do BLAST. A metodologia desenvolvida neste trabalho é apresentada em forma de algoritmo a seguir. A entrada do algoritmo é um conjunto (S1) de sequências nucleotídicas codificadoras de proteínas de uma família do *T. cruzi*.

```
Início do Algoritmo
   Fazer i = 1;
   Fazer u = falso;
   Enquanto u for igual a falso fazer:
      Executar o BLASTN para obter os alinhamentos de Si com Si;
      Filtrar os alinhamentos utilizando a estratégia de filtragem;
      Se i > 1
         Separar os alinhamentos filtrados em grupos.
         Se todos os grupos forem unitários fazer:
            u = verdadeiro;
            Interromper o loop;
         Senão
            Aplicar a estratégia de corte nos grupos gerando o conjunto de sequências Si+1;
      Senão
         Aplicar a estratégia de corte aos alinhamentos filtrados gerando o conjunto Si+1;
      Fazer i = i + 1;
   Fim do Enquanto;
   Executar o BLASTN para obter os alinhamentos de Si contra Si;
   Definir os módulos utilizando a estratégia de definição de módulos;
Fim do Algoritmo.
```

8. Resultados e Discussão

O algoritmo desenvolvido neste trabalho, descrito na Seção 4.3.3, foi implementado em C++ e executado para as sequências nucleotídicas codificadoras das proteínas da família MASP do *T. cruzi* com diferentes conjuntos de valores para os parâmetros do BLAST. Para uma melhor avaliação dos resultados obtidos, o alinhamento dos possíveis módulos com as sequências originais foi utilizado para mapear, para cada módulo, as posições em que ocorre nas sequências da família e para calcular sua frequência de ocorrência.

Posteriormente alinhou-se o conjunto de sequências originais com o conjunto de módulos, mapeando para cada sequência as posições dos módulos, que ela apresenta e calculando a frequência com que ocorrem. A Tabela 4 apresenta os resultados encontrados para cada combinação de parâmetros testada, onde os códigos dos parâmetros são *-e* para *E-value*, *-F* para filtro de regiões de baixa complexidade, *-r* para recompensa de igualdade de nucleotídeos, *-q* para penalidade de desigualdade de nucleotídeos e *-W* para tamanho de palavra.

Os conjuntos de valores de parâmetros testados foram:

- $\mathbf{C_1}$: -F T, -e 10, -W 11, -r 1, -q -3;
- $\mathbf{C_2}$: -F F, -e 10, -W 11, -r 1, -q -3;
- $\mathbf{C_3}$: -F F, -e 100, -W 11, -r 1, -q -3;
- $\mathbf{C_4}$: -F F, -e 500, -W 11, -r 1, -q -3;
- $\mathbf{C_5}$: -F F, -e 1000, -W 11, -r 1, -q -3;
- $\mathbf{C_6}$: -F F, -e 500, -W 10, -r 1, -q -3;
- $\mathbf{C_7}$: -F F, -e 500, -W 9, -r 1, -q -3;
- $\mathbf{C_8}$: -F F, -e 500, -W 8, -r 1, -q -3;
- $\mathbf{C_9}$: -F F, -e 500, -W 12, -r 1, -q -3;
- $\mathbf{C_{10}}$: -F F, -e 500, -W 13, -r 1, -q -3;
- $\mathbf{C_{11}}$: -F F, -e 500, -W 12, -r 1, -q -2;
- $\mathbf{C_{12}}$: -F F, -e 500, -W 12, -r 1, -q -1.

O valor sete indicado por [18] para tamanho inicial de palavra não consta entre os resultados apresentados, pois houve erro de execução do BLAST para todas as combinações de valores de parâmetro testadas envolvendo esse valor de tamanho de palavra. Isso se deve possivelmente à ausência de memória necessária à execução correta, devendo-se ressaltar que a configuração da máquina utilizada contava com processador de dois núcleos de 2GHz e 4GB de memória RAM. Na impossibilidade de execução do algoritmo com o valor indicado na literatura, testou-se o tamanho inicial de palavra com valores oito e nove. Os resultados com combinações desses valores se mostraram insatisfatórios visto que poucos módulos foram encontrados e o mapeamento desses resultou em várias sequências desprovidas de módulos.

Conjunto de Valores de Parâmetros	Total de Módulos	Média de Ocorrências	Média de Módulos por Sequência	Máximo de Módulos por Sequência
C_1	496	29	14	39
C_2	527	28	16	60
C_3	1300	17	24	97
C_4	1045	24	28	69
C_5	1043	24	28	67
C_6	530	21	12	39
C_7	169	25	4	20
C_8	24	32	0	7
C_9	**2464**	**16**	**42**	**183**
C_{10}	1235	17	23	147
C_{11}	2364	17	42	168
C_{12}	1368	21	30	136

Tabela 4 – Comparativo de Resultados de Combinações de Valores de Parâmetros do BLAST.

Para comparação de resultados foi atribuído maior peso ao número de módulos encontrados e às frequências de ocorrência (valores médios de ocorrência de módulos e de ocorrência de módulos por sequência), considerando-se melhores os maiores valores. A comparação de C_1 (valores padrão) com C_2 mostrou que a desativação do filtro de regiões de baixa complexidade levou a melhores resultados. Dessa forma todas as outras combinações de valores de parâmetros testados utilizam o filtro desativado. As comparações de C_2 com C_3 e de C_3 com C_4 mostram que a elevação do valor de *E-value* produziu melhores resultados. Comparando-se C_4 e C_5 pode-se observar a obtenção de resultados equivalentes. Assim, como o valor *-e 500* necessita de menos recursos de máquina para execução, as demais combinações testadas consideram esse valor. A comparação de C_4 com C_6, C_7 e C_8 mostra que a redução do valor inicial de palavra levou à produção de resultados piores que o uso do valor padrão; e com C_9 mostrou que o aumento do tamanho inicial de palavra melhorou os resultados. No entanto, a comparação de C_9 e C_{10} mostrou que a elevação de mais uma unidade no tamanho inicial de palavra provoca uma piora nos resultados; sendo assim, o tamanho inicial de palavra 12 foi utilizado para as demais combinações. Comparando-se C_9 com C_{11} e C_{12} observa-se uma piora nos resultados com C_{12} e uma leve piora nos resultados com C_{11}, o que indicou que o sistema de pontuação padrão obteve melhores resultados.

A análise dos resultados permitiu verificar que a combinação de valores de parâmetro C_9, que associa sistema de pontuação padrão para igualdade e desigualdade de nucleotídeos, alto valor de *E-value*, desativação do filtro de regiões de baixa complexidade e ligeira elevação em relação ao valor padrão do tamanho inicial de palavra, foi a que apresentou melhores resultados na execução da metodologia proposta para a família de proteínas MASP do *T. cruzi*.

O processamento da família MASP com o conjunto de valores C9 levou quatro iterações do algoritmo, distribuídas ao longo dos 192 minutos de execução utilizando uma máquina de 4GB de RAM com processador Core 2 Duo de 2GHz. Aplicada a estratégia de definição de módulos, foram definidos 2464 módulos, sendo que cada um ocorreu em média 16 vezes ao longo das sequências da família. A Tabela 5 apresenta os cinco módulos de maior incidência na família MASP.

Módulo	Número de Ocorrências
GTGGTGGCCGCGTGA	646
CGTGTGCTGCTGGTGTGTGCCCT CTGCGTG	616
CCCCTCTTTTGC	602
GCGATGATGATG	478
GGCGACAGTGAC	363

Tabela 5 – Módulos de Maior Incidência nas Proteínas da Família MASP

A Figura 5 apresenta uma visualização gráfica do mapeamento de módulos para a sequência de maior incidência desses (183 módulos). Na Figura é possível observar a ocorrência de sobreposição de módulos, o que acontece devido ao próprio BLAST relatar alinhamentos sobrepostos por alinhar uma mesma região da *query* mais de uma vez com a mesma sequência do banco de dados. A sobreposição sugere que os módulos obtidos ao fim do algoritmo não constituem necessariamente módulos individuais, havendo a possibilidade de serem combinados para formar outros módulos em um processo de refinamento dos resultados.

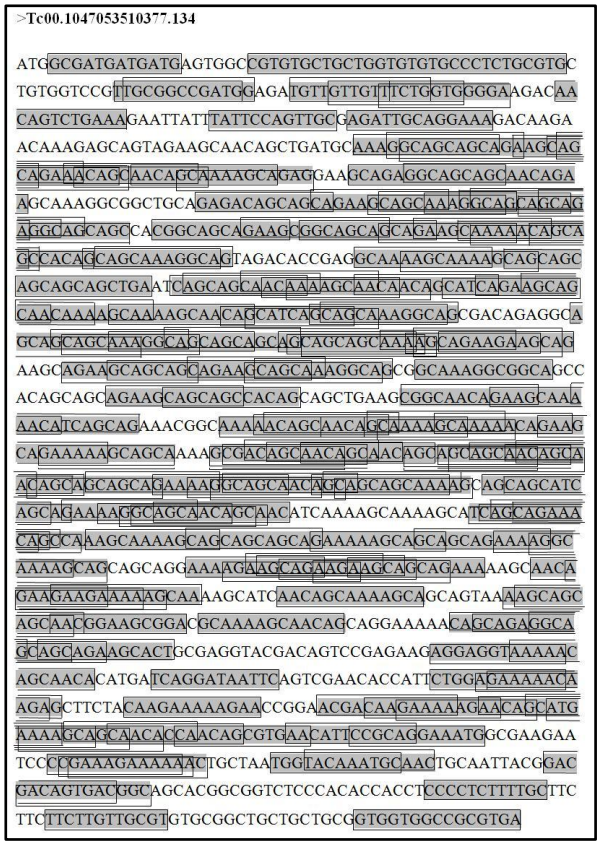

Figura 5 – Mapeamento de Módulos na Sequência Tc00.1047053510377.134

9. Conclusão

Este trabalho se propôs a criar uma metodologia para identificação de módulos formadores de sequências nucleotídicas codificadoras de proteínas mosaicas do *Trypanosoma cruzi* e constituintes do transcriptoma do parasito utilizando a ferramenta BLAST.

O algoritmo para a metodologia foi implementado e executado com diferentes combinações de parâmetros do BLAST a fim de comparação dos resultados obtidos. Como medidas de comparação, foram utilizados o número total de módulos encontrados e os valores médios de ocorrência de módulos e de ocorrência de módulos por sequência. Pela observação dos resultados se concluiu que a metodologia provou ser eficaz para identificação de módulos formadores de proteínas mosaicas a partir das sequências nucleotídicas que as codificam e que a combinação de desativação do filtro de regiões de baixa complexidade, alto valor de E-value, ligeira elevação do tamanho inicial de palavra em relação ao valor padrão e sistema de pontuação de nucleotídeos padrão apresentou os melhores resultados para a família de proteínas MASP do *T. cruzi*.

A partir dos resultados obtidos se concluiu também que foi confirmada a estrutura mosaica das proteínas da família MASP, visto que o mapeamento dos módulos encontrados possibilitou a visualização desses em todas as sequências da família com uma média de 42 módulos por sequência.

É proposto como trabalho futuro a comparação dos resultados obtidos neste trabalho com os encontrados por [19], cuja identificação de módulos de proteínas mosaicas do *T. cruzi* se baseia no proteoma do parasito.

Como foi observada a sobreposição e ocorrência em série de alguns módulos, é proposto como trabalho futuro ainda o estudo da ocorrência condicional de módulos, ou seja, da possibilidade de ocorrência de um módulo estar condicionada à ocorrência de outro, o que possibilitaria o refinamento dos resultados obtidos neste trabalho por meio da redefinição como módulo único de módulos que se sobrepõem ou que ocorrem sempre em série. Além disso, o estudo da ocorrência condicional de módulos e da presença de um mesmo conjunto de módulos em diferentes sequências pode trazer informações importantes para estudiosos do *T. cruzi* e da Doença de Chagas.

Referencial Bibliográfico

[1] BROWN, T. A. **Genomes**. 2. ed. Oxford: BIOS Scientific Publishers, 2002. 572 p.

[2] NHGRI – NATIONAL HUMAN GENOME RESEARCH INSTITUTE. **Transcriptome**. 2008. Disponível em: <http://www.genome.gov/13014330>. Acesso em: 06 maio 2008.

[3] AVERY, V. M.; ADRIAN, D. L.; GORDON, D. L. Detection of mosaic protein mRNA in human astrocytes, **Immunology and Cell Biology**, v. 71, n. 3, p. 215-219, june 1993.

[4] GABORIAUD, C.; ROSSI, V.; FONTECILLA-CAMPS, J. C.; ARLAUD, G. J. Evolutionary Conserved Rigid Module-domain Interactions can be Detected at the Sequence Level: The Examples of Complement and Blood Coagulation Proteases. **Journal of Molecular Biology**, v. 282, n. 2, p. 459-470, sep 1998.

[5] HEGYI, H.; BORK, P. On the classification and evolution of protein modules. **Journal of Protein Chemistry**, v. 16, n. 5, p. 545-551, july 1997.

[6] KOLKMAN, J. A.; STEMMER, W. P. C. Directed evolution of proteins by exon shuffling. **Nature Biotechnology**, v. 19, n. 5, p. 423-428, may 2001.

[7] DOOLITTLE, R.F. The multiplicity of domains in proteins. **Annual Review of Biochemistry**, v. 64, p. 287-314, july 1995.

[8] PATTHY, L. Modular exchange principles in proteins. **Current Opinions in Structural Biology**, v. 1, p. 351-361, 1991.

[9] ANDRADE, L. O.; ANDREWS, N. W. The *Trypanosoma cruzi* host cell interplay: location, invasion, retention. **Nature Reviews Microbiology**, v. 3, n. 10, p. 819-823, oct. 2005.

[10] FRASCH, A. A. C. Functional diversity in the trans-sialidase and mucin families in *Trypanosoma cruzi*. **Parasitology Today**, v. 16, n. 7, p. 282-286, july 2000.

[11] KAHN, S. J.; NGUYEN D.; NORSEN, J.; WLEKLINSKI, M.; GRANSTON, T.; KAHN, M. *Trypanosoma cruzi*: monoclonal antibodies to the surface glycoprotein superfamily differentiate subsets of the 85-kDa surface glycoproteins and confirm simultaneous expression of variant 85-kDa surface glycoproteins. **Experimental Parasitology**, v. 92, n. 1, p. 48-56, may 1999.

[12] DEGRAVE, W. ***Trypanosoma cruzi*:** o genoma. Rio de Janeiro. Disponível em: <http://www.fiocruz.br/chagas/cgi/cgilua.exe/sys/start.htm?sid=14>. Acesso em: 05 maio 2008.

[13] EL-SAYED N.M.; MYLER, P.J.; BARTHOLOMEU, D.C.; NILSSON, D.; AGGARWAL, G.; TRAN, A.N.; GHEDIN, E.; WORTHEY, E.A.; DELCHER, A.L.; BLANDIN, G.; WESTENBERGER, S.J.; CALER, E.; CERQUEIRA, G.C.; BRANCHE, C.; HAAS, B.; ANUPAMA, A.; ARNER, E.; ASLUND, L.; ATTIPOE, P.; BONTEMPI, E.; BRINGAUD, F.; BURTON, P.; CADAG, E.; CAMPBELL, D.A.; CARRINGTON, M.; CRABTREE, J.; DARBAN, H.; DA SILVEIRA, J.F.; DE JONG, P.; EDWARDS, K.; ENGLUND, P.T.; FAZELINA, G.; FELDBLYUM, T.; FERELLA, M.; FRASCH, A.C.; GULL, K.; HORN, D.; HOU, L.; HUANG, Y.; KINDLUND, E.; KLINGBEIL, M.; KLUGE, S.; KOO, H.; LACERDA, D.; LEVIN, M.J.; LORENZI, H.; LOUIE, T.; MACHADO, C.R.; MCCULLOCH, R.; MCKENNA, A.; MIZUNO, Y.; MOTTRAM, J.C.; NELSON, S.; OCHAYA, S.; OSOEGAWA, K.; PAI, G.; PARSONS, M.; PENTONY, M.; PETTERSSON, U.; POP, M.; RAMIREZ, J.L.; RINTA, J.; ROBERTSON, L.; SALZBERG, S.L.; SANCHEZ, D.O.; SEYLER, A.; SHARMA, R.; SHETTY, J.; SIMPSON, A.J.; SISK, E.; TAMMI, M.T.; TARLETON, R.; TEIXEIRA, S.; VAN AKEN, S.; VOGT, C.; WARD, P.N.; WICKSTEAD, B.; WORTMAN, J.; WHITE, O.; FRASER, C.M.; STUART, K.D.; ANDERSSON, B. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. **Science**, v. 309, n. 5733, p. 409-415, july 2005.

[14] MOUNT, David W. **Bioinformatics:** sequence and genome analysis. 2. ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2004. 692 p.

[15] KORF, I.; YANDELL, M.; BEDELL, J. **BLAST:** An essential guide to the Basic Local Alignment Search Tool. Sebastopol: O'Reilly, 2003. 339 p.

[16] MAYER, H. **A collection of evaluated bioinformatics programs and databases:** Sequence Similarity. Disponível em <http://homepage.univie.ac.at/herbert.mayer/>. Acesso em 11 ago. 2008.

[17] INCOGEN. **NCBI Blastn**. Disponível em <http://www.incogen.com/public_documents/vibe/details/NcbiBlastn.html>. Acesso em: 13 ago. 2008.

[18] NCBI – NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Basic Local Alignment Search Tool**. Disponível em: <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE_TYPE=BlastFAQs>. Acesso em : 13 ago. 2008.

[19] GOMES, A. S.; SOUZA, T. R. **Uma Metodologia para Identificação de Módulos Formadores de Sequências de Proteínas Mosaicas do Trypanosoma cruzi a partir do Proteoma do Parasito Utilizando a Ferramenta BLAST**. 2008. 47p. Monografia (Graduação em Ciência da Computação) – Universidade Federal de Lavras, Lavras, MG.